ROYAL NETHERLANDS ACADEMY OF ARTS AND SCIENCES

THE COMPUTATIONAL HUMANITIES PROGRAM

THE COMPUTATIONAL HUMANITIES PROGRAM IS FOCUSED AT THE ENHANCEMENT OF TECHNOLOGY IN THE HUMANITIES

Grant application form

1) Title: Census data open linked – CEDA_R
From fragment to fabric – Dutch census data in a web of global cultural and historic information

2) Main applicant(s):

Dr. Peter Doorn DANS
Professor Erik-Jan Zürcher IISG
Professor Frank van Harmelen VU

3) Summary (max 250 words)

Currently, Europe is confronted with industrial restructuring, migration, aging of population and financial crisis in a world of accelerated change. Learning from (social-economic) history helps to understand the interrelation between macro-economic change and individual lifestyles, policy regimes, labour markets, communities and national wealth. However, sources of historical information about the lives of individuals, communities, and nations are still scattered.
This project takes Dutch census data as a starting point to build a semantic data-web of historical information. With such a web we will answer questions such as:

- What kind of patterns can we identify and interpret in expressions of regional identity?
- How to relate patterns of changes in skills and labour to technological progress and patterns of geographical migration?
- How to trace changes of local and national policies in the structure of communities and individual lives?

Census data alone are not sufficient to answer these questions. This project applies a specific web-based data-model – exploiting the Resource Description Framework (RDF) technology– to make census data inter-linkable with other hubs of historical socio-economic and demographic data and beyond. Pattern recognition appears on two levels: first to enable the integration of hitherto isolated datasets, and second to apply integrated querying and analysis across this new, enriched information space. Data analysis interfaces, visual inventories of historical data and reports on open-linked data strategies for digital collections are results of this project. The project will result in generic methods and tools to weave historical and socio-economic datasets into an interlinked semantic data-web.

4) Research group

Name	Institute	Fte	Role/Competence	Costs
Dr. Andrea Scharnhorst	DANS/e-hum	0.2 fte/week	Project coordination, Complexity science, Information science	Not applicable[1]
N.N. PhD Linked data curation in social history using semantic web	DANS	1fte/4 years	Computer sciences, History	203.692 EUR
N.N. PhD Data models of Dutch census data	IISG	1fte/4 years	History, Computer sciences	203.692 EUR
N.N. PostDoc	DANS	1fte/4 years	Complexity science	264.248 EUR
Prof. Kees Mandemakers	IISG Erasmus School of History, Culture and Communication, Erasmus University Rotterdam	0.02 fte /w	Data Harmonization Large Historical Databases Historical Sample of the Netherlands (HSN) Promotor PhD Data models	Not applicable
Professor Jan Kok	IISG Radboud University Nijmegen	0.02 fte /w	CLIO-INFRA/HSN	Not applicable
Dr. O. Boonstra	Radboud University Nijmegen	0.02 fte/w	Historical Informatics Co-promotor	Not applicable
Professor Frank van Harmelen	VU, Faculty of Sciences, Knowledge Representation and Reasoning Group	0.02 fte/w	Computer Sciences, Knowledge representation and reasoning group Promotor PhD Linked data curation using semantic web (co-promotor for the PhD Social History)[2]	Not applicable
Dr. Rinke Hoekstra	VU, Faculty of Sciences	0.02 fte/w	RDF models	Not applicable
Dr. Dirk Roorda	DANS	0.02 fte/w	R&D Implementation	Not applicable
Maarten Hoogerwerf	DANS	0.02 fte/w	OpenLinkedData implementation	Not applicable
Dr. Rene van Horik	DANS	0.02 fte/w	Data curating practices, History	Not applicable

[1] Staff at the KNAW institutes and universities will contribute to the project relying on their basic funding. No costs are budgeted to be covered by the CompHum programme.
[2] This function might be filled by another member of the group of Frank van Harmelen

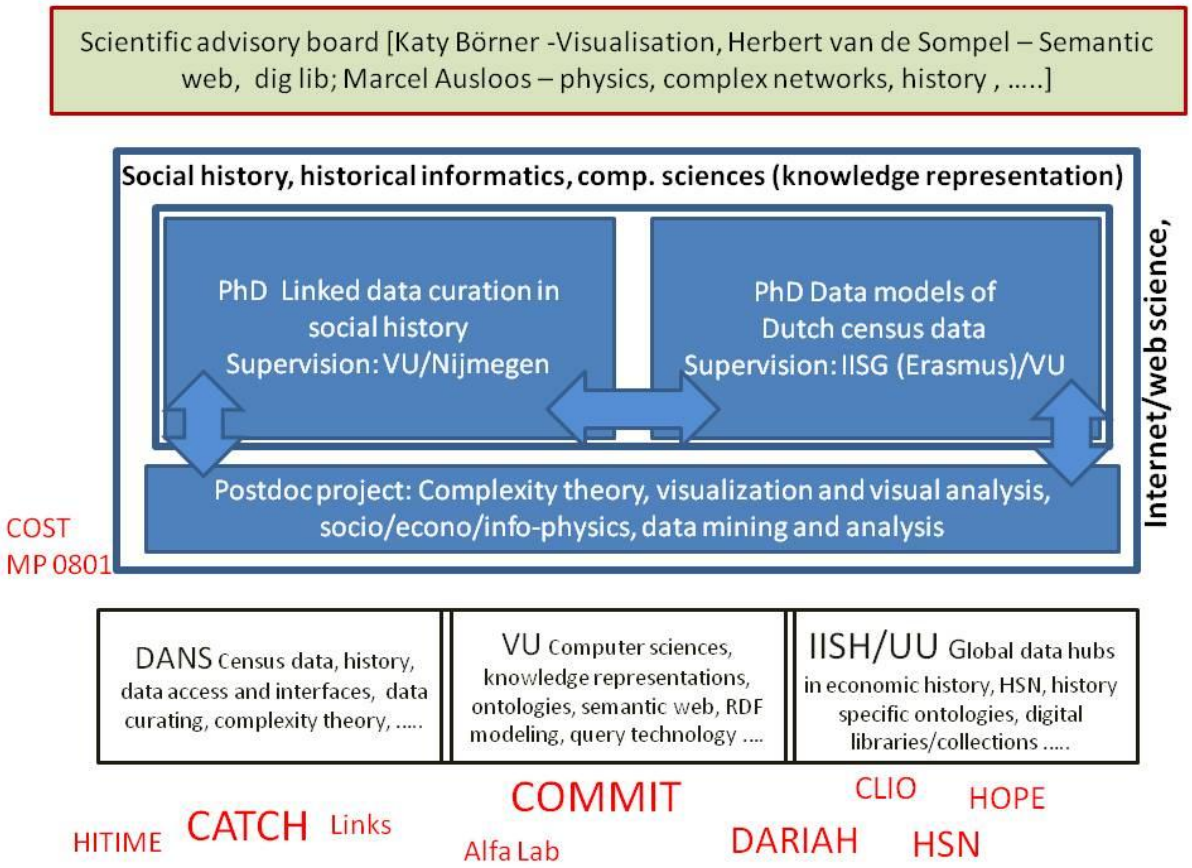


Figure 1: Project structure, competences and related projects

Composition of the research team/integration of subprojects: This project builds on two PhD sub-projects (running in parallel) which will be supervised by a team of computer scientists (VU), social historians (Erasmus, Radboud) and historical informatics specialists (Radboud). The coupled PhD strategy establishes new innovative collaboration around social history informatics, and in particular extensions towards research fronts in computer and information sciences, while, at the same time, informing them about special challenges coming from humanities research questions. The third – interrelating - sub-project (a postdoc) establishes a link to web science or internet science in a broader sense. The integration between the subprojects will be ensured by combined supervision of the PhD's, a (temporal) co-location of the researchers at participating institutes and the integration into the new e-humanities group. The scientific advisory board is formed by the main applicants, and international experts.

Unlocking other funding: The project will be closely aligned to the newly started nationally funded COMMIT programme. We will exploit and build on tools and methods developed in COMMIT subproject P23 ("from data to semantics for data-publishers"). Generic tools for data-modelling, data-publishing, query-construction and provenance tracking will be adopted for the specific requirements from Computational Humanities research. In turn, COMMIT P23 has committed to using the census data as one of their key case-studies in publishing linked scientific datasets (besides data from life-sciences and healthcare). This effectively unlocks the 1.8M€ COMMIT P23 additional budget to support the purposes of Computational Humanities research.

Leveraging existing research: The research group uses existing contacts with Herbert van de Sompel [Los Alamos, USA] (currently fellow of the KNAW Visiting professor programme, located at DANS), Katy Börner [Indiana University, USA] (fellow of the KNAW Visiting professor programme at DANS/e-hum 2012). Links will be established to other on-going research projects at the KNAW as CLIO-INFRA (Jan Luiten van Zanden, Jan Kok) and the European Historical Population Sample Network (EHPS, Kees Mandemakers) concerning other data hubs and visualization, and the

Circulation of Knowledge Project (Charles van den Heuvel) concerning concept identification and meta data visualization. The current COST action MN 0801 "Physics of Competition and Conflicts" (until 2012) (Andrea Scharnhorst head of WG 1 "Information and Knowledge", Marcel Ausloos MC member) is relevant concerning complex networks analysis. The project relies on results of the current *Strategiefonds* project "Knowledge Space Lab" (led by Andrea Scharnhorst) concerning visual representations of knowledge orders and their evolution over time. Existing links to the Institute for Library Science at the Humboldt University Berlin (DFG Project "Research diversity", Andrea Scharnhorst, advisory board) will be extended to other initiatives of the institute relevant for this project, in particular concerning the Europeana project (Stefan Gradman). In this context, the HOPE project at the IISG is also of relevance.

5) Innovative value (max 500 words)

The challenge

Jigsaw pieces of historical information can be found at different places (archives, statistics bureaus, museums, scientific institutes). Despite the tremendous efforts in digitization and web-based technologies, data is still notoriously hard to find, access, interpret, share, and reuse. The complexity and abundance of data resources require strategies for an innovative knowledge and metadata management.

As statistical information, census data have a special place in our historical memory, located at a meso-level between on the one hand the traces left by individuals in the population administration and ego-documents, and on the other hand the higher aggregated statistical information in national and international statistics. Census data also form an essential building block of current efforts to realize global data hubs with historical information on population size and composition, migration rates, labour relations, education, and economic performance (see www.clio-infra.eu for an example effort in this direction).

Although easy access to a historical data web is of interest for everyone, including the public, the specialist communities need very fine-grained standardized and harmonized databases, which allow to download or remotely query data for research. Changing categorizations form a problem: for general retrieval as well as for standardized, long-term statistical analysis. This holds in particular for incomplete, differently coded statistical historical information.

The innovative value op the project

The project analyses, compares and satisfies these different needs by combining two PhD theses and a postdoc project which bridges between the PhD projects by addressing generic, structural questions from the perspective of complex social systems.

The project combines current computational history with efforts of the cross-disciplinary, web-based knowledge representations (semantic web).

The project contributes to pattern recognition and formalization applied to historical data in the following two ways: (1) patterns will be formulated as queries over distributed, heterogeneous, semi-structured interlinked databases, turning pattern recognition into query-answering over such datasets; (2) techniques from complex systems and network analysis will be used to detect historical changes and to differentiate between modest and more radical changes in society.

The project aims at a better visibility and searchability of historical data sources for the wider public as well as better access and interchangeability of data for specialists.

The project will integrate the Dutch census data into the semantic web by using an RDF data model. This is a first for large-scale socio-economic historical data.

Key enabler: Most if not all Computational Humanities research crucially depends on the availability of datasets. This project paves the way towards an innovative machine readable, distributed and accessible archive for humanities datasets in the linked open data cloud. The tools and methods generated as spin-off from this project will enable many other Computational Humanities research efforts to much more effectively publish and use their datasets.

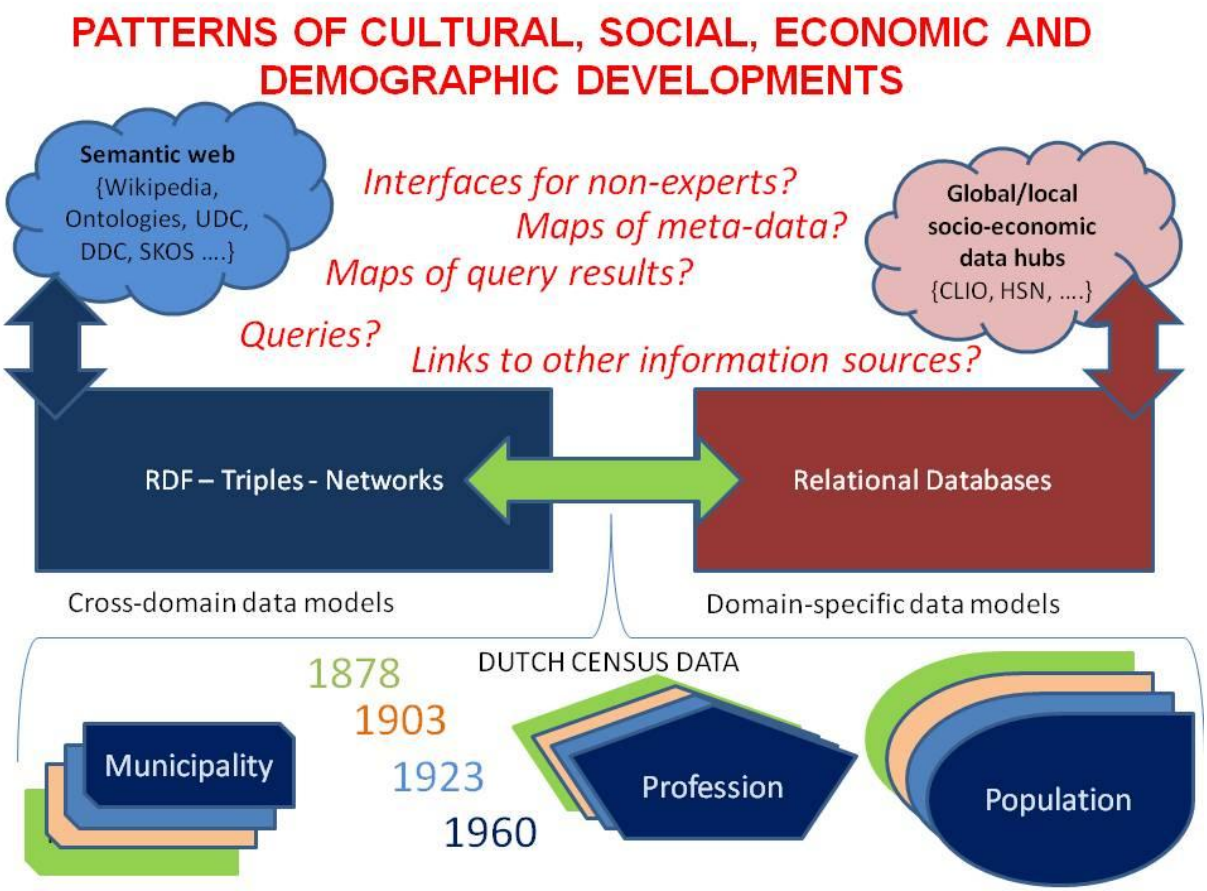


Figure 2: Scope and challenges in the project - Data modelling inside of a knowledge domain concentrates on establishing links between different temporal slices of historical data. Data modelling across knowledge domains as envisioned in semantic web technologies aims at linking between different sources of knowledge.

6) Description of Research Plan (max. 3 pages)

State of the art of the national census collection

The collection of the Dutch census (1795-present) consists, among other resources, of thousands of data tables (currently as spreadsheets). The dataset contains information on demographic, social, economic and cultural characteristics of the Dutch population over the past two centuries, described at the level of neighbourhoods, municipalities, regions, provinces, cities and the rural areas. DANS has created a digital interface to the Dutch census data on the website www.volkstellingen.nl which allows a simple search and query across different census data and attributes. Other projects have provided fine grained access to the data, but only for special segments. For example, the HASH (Hub for Aggregated Social History) service facilitates querying historical data by municipality in various heterogeneous sources. The CBS has made an interactive version of the 1899 and 1930 census available. To summarize, there exists no integrated access to this Dutch census data. Moreover, each of the existing spreadsheets uses its own set of categories. This makes cross-analysis between different census sets very time-consuming. One urgent need of the specialist communities is to create a homogeneous access to the Dutch census data comparable to other international projects (e.g., the Integrated Public Use Microdata Series of the Minesota Population Center (see http://www.ipums.org/...).

As said before, statistical information census data have a special place in our historical memory, located at a meso-level between the traces left by individuals in the population administration, but also in ego-documents stored in museums and archives, in collections of individual life courses reconstructed in projects such as the Historical Sample of the Netherlands (HSN), and the higher aggregated statistical information at national and international level. This brokering position of census data makes the data particular suitable as a crystallization point for an emerging semantic historical data-web. Suddenly, links to holdings of museums, libraries and archives are possible. At the same time, in social history, census data form an essential building block of currents efforts to build global data hubs with historical information on population size and composition, migration rates, labor relation, education, and economic performance (see http://www.clio-infra.eu). Moreover, for databases with longitudinal information on the micro-level, such as the HSN, or other databases within the EHPS-network, census-data provide a higher-level context relevant analysis of micro-data providing data. In any case, to be useful, census data need first to be modelled, i.e. classified and coded in a well structured way. In this project, different data models will be applied and compared. The HSN-database will be used to test RDF-coding of census data in combination with other kind of databases.

State of the art in semantic web technology

Graph-based ("noSQL") data-models are a recent and promising approach to providing homogeneous access to physically distributed and semantically heterogeneous data. In this project we focus on the so-called Resource Description Framework, which exploits web-technology (URIs, HTTP) to deal with physical distribution and knowledge representation (ontologies) for semantic heterogeneity. RDF data modelling can be described as creating a shell around data sets. This shell leaves the original data structure intact, but makes it interoperable with other data sources for which also RDF shells exist. The RDF data model has been successful in providing interoperability to scientific datasets in the life-sciences, chemistry and other natural sciences. The Dutch census dataset is ideally suited to test the validity of the RDF data model in the case of humanities research.

A further advantage of RDF is that attributes do not need to be mapped to each other a priori (as in relational databases)– but that their multiple relations can be added as specific structures in the data graph. Changing ontologies over time become visible in changing data graph structures over time. However, a computational challenge is to create user interfaces and API's (Advanced Programming Interfaces) for RDF graphs. Another challenge is to create bridges between different RDF islands of knowledge domains. Generic attributes such as geographic references, specific events in time, and actors (names) can be used to correlate RDF spaces. With this correlation, specific knowledge units are immediately embedded in a self-organizing, emerging semantic web. Based on the Open Data movement, existing royalty-free datasets (such as Wikipedia, Geonames, Wordnet, and dozens of

others) become increasingly available as RDF. Interlinking them on a large scale allows not only to retrieve where information can be found, but also to relate different types of information to each other. RDF modelling in the area of census data (American Factfinder Project) so far has primarily been used in the context of making public data easier available and creating transparency for governmental information. In this project, we first apply the RDF approach for social-historical research.
The CEDA_R project responds

- to the need for a harmonized and integrated database of census data
- the efforts to create interlinked data-hubs such as CLIO-INFRA
- to efforts to integrate historical information in a semantic data-web

Main objectives of the project

Objective 1: Co-develop database models and RDF models for the set of Dutch census data
Objective 2: Embed Dutch census data in the emerging semantic data web
Objective 3: Develop visual enhanced retrieval principles

Objectives in more detail and allocated work packages

O1. Co-develop database models and RDF models for the set of Dutch census data (WP1, WP 2)

1.1. Compare different data model strategies on their appropriateness in the field of social history and in relation to a cross-disciplinary use and access to historical data
1.2. Analyze the relevance of data strategies for content-related questions such as the kind of patterns to be identified in the enormous social, economic and demographic changes that took place during the 19th and 20th century. We consider here changes in skills and labour in relation to technological progress and patterns of changing household composition and geographical migration.
1.3. Enable meta-data mapping (which historic sources are available with which quality in which period of time) across different data sources.

O2. Embed Dutch census data in the emerging semantic data web. (WP1,2,3)

2.1. Construct links to (historical) geographic knowledge, academic resources about census data and the corresponding historical period in general, and holdings of other archives and museums. Preferably by exploiting (semi-)automated semantic mapping technologies.
2.2. Use RDF described census data to map geographic-political organization over time, the development of the labour world, and changes of family structures.
2.3. Treat RDF data graphs as complex evolving networks and study their structural and temporal properties.

O3. Develop visual enhanced retrieval principles. (WP 4,1,2)

3.1. A (geographic) map on which available data sources can be displayed
3.2. A network of data sources together with a tree-map display of the size and structure of collections such as Dutch census
3.3. An interface in which simple statistical information can be retrieved.

Description of work[3]

Workpackage 1: Dutch census data in the semantic web - RDF data modelling

This workpackage applies RDF strategies to Dutch census data. Using RDF to access the central dataset (which basically resembles a world-wide census) allows testing the capabilities and limits of RDF, without having to deal at the same time with the harmonization and ambiguity problems involved in the Dutch censuses. In the project we apply the "Minimality principle" as applied by the Europeana project (making use of elements from already defined WWW namespaces wherever possible and define our own elements preferably as specialisations of existing ones and only if strictly required). We expect that a high ambiguity (multiple links, changing network structure) will indicate interesting changes in the history of society. By developing an RDF model for Dutch census data, by

[3] Collaboration between the different researchers in the project for the different tasks and the timeline are visualized in Figure 3.

relating this to RDF models of other census data, and by linking this to RDF representations of other relevant vocabularies we will look for effects of synchronization or shifts in changes on a more global scale.
T1.1. RDF model & links to other semantic web sources; review of existing data models of census data, adaptation and modification, construction of the RDF model
T1.2. Query design (specific to different user communities)
T1.3. Best practice report to enable take-up of linking and re-use of data in other scientific disciplines and take-up in other KNAW institutes.
T1.4. Visual navigation through RDF modelled information spaces
T1.5. PhD thesis - **Linked Data curation in social history**

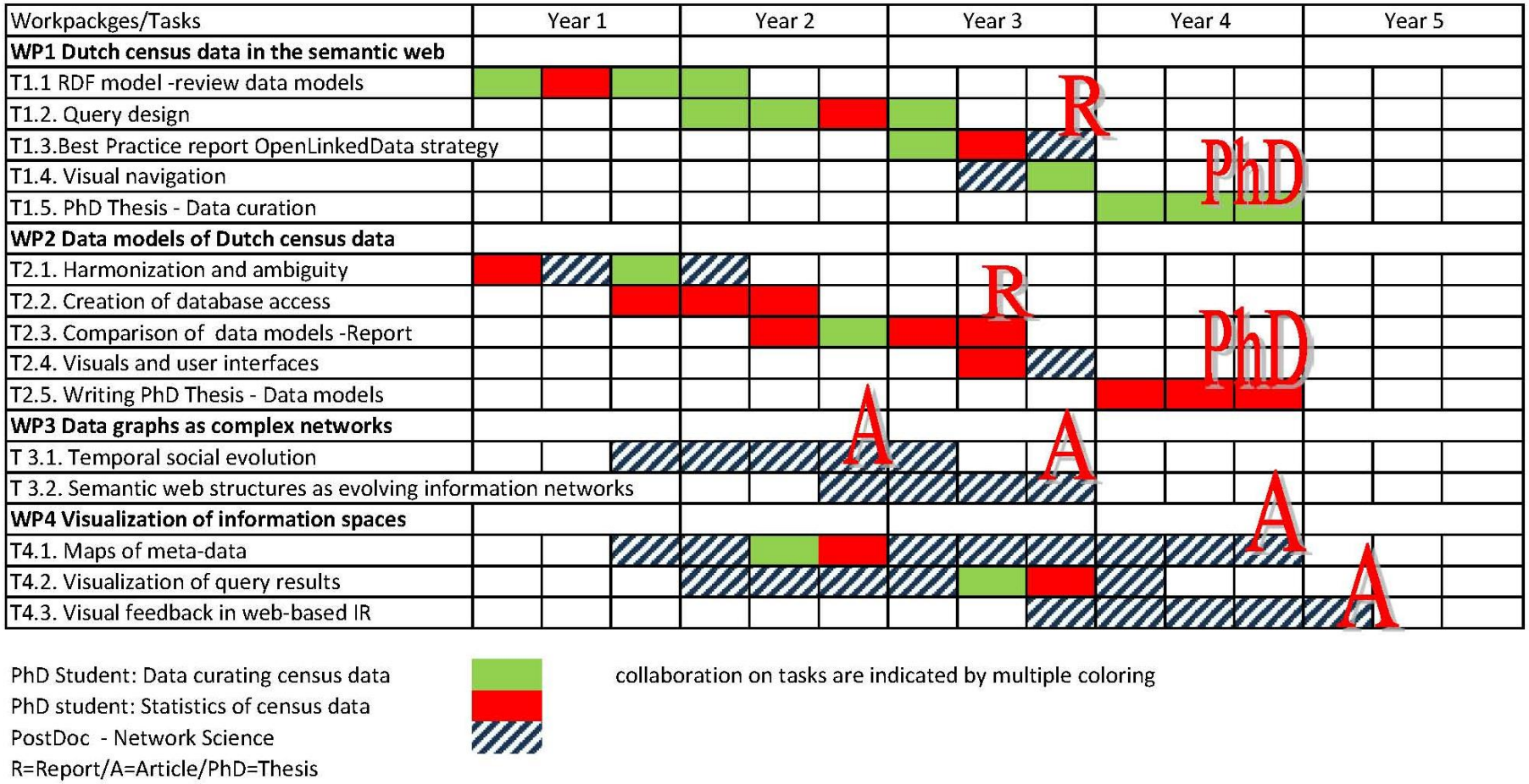
Workpackage 2: Data models of Dutch census data – designing an integrative database
Before starting RDF-coding it is necessary to process these data into one database. A first form of harmonization is more or less a semi-automatic result of this processing. In the building of an integrated and harmonized database from existing spreadsheets we make use of (peer) expertise in the CLIO-INFRA project concerning issues of harmonization, data ambiguity, and standardization. In a second stage, existing classifications in the census will be standardized and the data will be further harmonized and related to an RDF-coding. Important will be the evaluation of the integration of the census-ontology into the more general ontologies. Another aspect is the (visual) analysis of temporal changes and ambiguities in data modelling systems.
T2.1. Harmonization and ambiguity of the existing census data
T2.2. Creation of a database access and standard retrieval interface- user community specific: new web interface to Dutch census data. Test of the interface with different user communities (e.g., master class of history students, e-hum workshop on tools for humanities scholars)
T2.3. Comparison of different data models: Connect the RDF modelling with the creation of a harmonized relational data base - Report
T2.4. Visual elements in user interfaces to socio-historical, statistical data
T2.5. PhD Thesis – **Data models and interfaces for a statistical analysis of historical census data**

Workpackage 3: Data graphs as complex networks
The structural properties of time-dependent large RDF data graphs can be analyzed from a network perspective. From these insights we will be able to better understand for which kind of questions the fine-granularity of domain-specific ontologies is needed and when we need rougher, bird-eye-view ordering of information. In particular, we ask: How does ambiguity of links (being nested, belonging to different ontologies, multiple links between multiple nodes, coupled networks) in the data graph reflect ambiguity of term matching? Can we gain feedback from automatic mapping and matching exercises to question current paradigms of defining epochs in history and related mapping of shifting attributes? In other words: Can experts learn or get inspired from the mistakes of machines? Can we apply techniques of community detection and block modelling to data graphs in order to reveal hidden patterns of similarity in massive data?
T3.1. Temporal social evolution mapped to data graphs: By analyzing the statistics of changes in terminology and attributes for which data are collected, we aim to differentiate between modest and more radical changes in society.
T3.2. Semantic web structures as evolving information networks: Highlight and analyze category changes as indicators of change in the sense of instabilities leading to new structure formation

Workpackage 4: Visual elements in the navigation through information spaces
Explore visualizations of the existing databases (using existing tools such as MagnaView). Visualize the emerging RDF graph (using existing tools such as Gephi). Visualize the larger environment of the census RDF in the semantic web. Develop visuals as additions to querying and browsing; address questions such as how many sources exist for a given query? How are query-results situated in the data-landscape?
T4.1. Maps of meta-data
T4.2. Visualization of query results
T4.3. Visual feedback in web-based information retrieval

Appendix

Figure 3: Time planning/Allocation of workload/Deliverables (Articles/PhD thesis, Reports)



7) Requested budget

Personell	Year 1	Year 2	Year 3	Year 4	Year 5	
PhD student Data Curation	50.923	50.923	50.923	50.923		
PhD student: Social history	50.923	50.923	50.923	50.923		
Postdoc	41.796	64.427	66.367	68.291	23.368	
Subtotal	*143.642*	*166.273*	*168.213*	*170.137*	*23.368*	
Material costs						
Durable equipment	2000	500	500	500		
Travel	1000	1500	1500	1000	800	
Other (experts, enganced publication)	1000	2000	2000	2000	2000	
Subtotal	*4000*	*4000*	*4000*	*3500*	*2800*	
Total	*147.642*	*170.273*	*172.213*	*173.637*	*26.168*	***689931,7***

8) Own investments

DANS started to enrich data/objects in its archives with the aim to create synergetic links between different collections and to foster their (re)-usability.
IISG is involved in different initiatives to harmonize databases with the same goal, but also based on recent achievements in information visualization such as network visualization, self-organized maps of multi-dimensional attribute spaces, and visual interfaces to databases.
In both institutions senior researchers with a background in socio-economic history and history computing are present which will provide a rich and perspective-heterogeneous research environment for this project. There exist also different connections to universities and domain-specific research schools such as Posthumus. Senior research staff at all participating institutes and universities allocate time to the project based on own funding (implicit matching).

In *Alfalab* (another initiative of the KNAW for innovative methods in social sciences and humanities) both institutes have worked together and first tests on OpenLinkedData are part of a Demonstrator. CEDA_R incorporates these results and experiences.

9) Continuation/implementation after ending the project

Dutch Census Data are one of the core data sets of DANS, the same holds for the HSN, and other databases available at IISG. From the project we expect steps towards interlinking them, relating them to other globally emerging data-hubs of this kind of data.

But, CEDA_R is more as one (statistical)data-driven project. With its orientation towards the semantic web and generic principles of knowledge organization it contributes to the emergence of a network of knowledge sources about history, located at libraries, museums and in specific collections.

KNAW institutes entail large collections of digitised objects of different natures, together with different systems of metadata. DANS started to enrich data/objects in its archives with the aim to create synergetic links between different collections and to foster their (re)-usability. This project contributes to processes of creation of meta-data, their mutual mapping and the building of intelligent information retrieval systems by means of visualizations.

In particular, The project delivers two reports (OpenDataLinks Strategies (in WP1) and Comparison of data-models (in WP 2)), pilots of simple but generic visual elements for data analysis, data retrieval and meta-data display (WP 4), and a new user interface to the Dutch Census Data (WP1+2).